

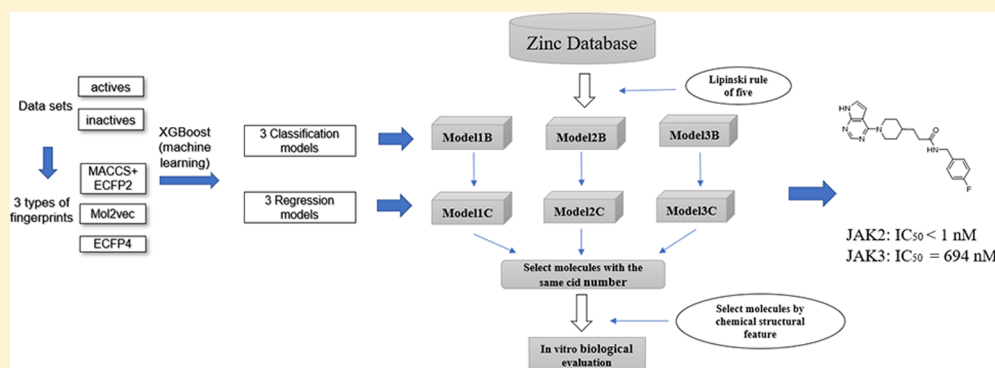
# Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors

Minjian Yang,<sup>†,‡</sup> Bingzhong Tao,<sup>‡</sup> Chengjuan Chen,<sup>†</sup> Wenqiang Jia,<sup>†</sup> Shaolei Sun,<sup>‡</sup> Tiantai Zhang,<sup>†</sup> and Xiaojian Wang<sup>\*,†,‡,§</sup>

<sup>†</sup>State Key Laboratory of Bioactive Substances and Functions of Natural Medicines, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100050, P.R. China

<sup>‡</sup>Joint Laboratory of Artificial Intelligence of the Institute of Materia Medica and Yuan Qi Zhi Yao, Beijing 100050, P.R. China

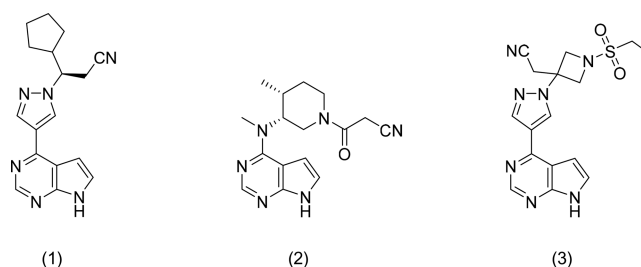
## Supporting Information



**ABSTRACT:** Developing Janus kinase 2 (JAK2) inhibitors has become a significant focus for small-molecule drug discovery programs in recent years because the inhibition of JAK2 may be an effective approach for the treatment of myeloproliferative neoplasm. Here, based on three different types of fingerprints and Extreme Gradient Boosting (XGBoost) methods, we developed three groups of models in that each group contained a classification model and a regression model to accurately acquire highly potent JAK2 kinase inhibitors from the ZINC database. The three classification models resulted in Matthews correlation coefficients of 0.97, 0.94, and 0.97. Docking methods including Glide and AutoDock Vina were employed to evaluate the virtual screening effectiveness of our classification models. The  $R^2$  of three regression models were 0.80, 0.78, and 0.80. Finally, 13 compounds were biologically evaluated, and the results showed that the  $IC_{50}$  values of six compounds were identified to be less than 100 nM. Among them, compound 9 showed high activity and selectivity in that its  $IC_{50}$  value was less than 1 nM against JAK2 while 694 nM against JAK3. The strategy developed may be generally applicable in ligand-based virtual screening campaigns.

## 1. INTRODUCTION

The Janus kinases (JAKs) are a family of intracellular non-receptor protein tyrosine kinases that play prominent roles in the cytokine-mediated JAK–STAT signaling pathway.<sup>1,2</sup> The JAK family consists of four enzymes, JAK1, JAK2, JAK3, and TYK2. To date, three drugs targeting JAKs have been approved by the U.S. Food and Drug Administration (FDA) (Figure 1). Ruxolitinib (1), a JAK1/JAK2 inhibitor, is approved for the treatment of primary myelofibrosis (PMF) in 2011.<sup>3,4</sup> Tofacitinib (2), a JAK1/JAK3 inhibitor with moderate activity on JAK2, and baricitinib (3), which inhibits JAK1 and JAK2, are approved for the treatment of rheumatoid arthritis (RA).<sup>5,6</sup> Among the four JAK subtypes, JAK2 emerged in the recent years as a potential therapeutic target for myeloproliferative neoplasm (MPN), which include polycythemia vera (PV), essential thrombocythemia (ET), and



**Figure 1.** Ruxolitinib (1), tofacitinib (2), and baricitinib (3).

PMF.<sup>7–10</sup> JAK2 inhibitors have been developed and evaluated in clinical trials for the treatment of MPN.<sup>11</sup>

**Received:** September 15, 2019

**Published:** November 20, 2019

Traditional ligand- and structure-based virtual screening (VS) approaches have been used in the discovery of JAK2 inhibitors. For instance, Jasuja and colleagues designed dual inhibitors of JAK2 and JAK3 by a pharmacophore- and docking-based VS approach.<sup>12</sup> Pharmacophore filtering and a three-dimensional quantitative structure–activity relationship (3D-QSAR) were used by Dhanachandra Singh and colleagues in the discovery of JAK2 inhibitors.<sup>13</sup> Over the past decade, machine learning (ML) algorithms, such as Support Vector Machine (SVM),<sup>14</sup> Random Forest (RF),<sup>15</sup> and deep learning-based methods,<sup>16</sup> have become increasingly popular for VS. SVM is a learning machine for a two-group classification problem. An SVM model developed by Liew and colleagues was able to identify novel Lck inhibitors and distinguish inhibitors from structurally similar noninhibitors at a false positive rate of 0.27%.<sup>17</sup> As a classification and regression tool, RF is introduced and investigated for predicting the quantitative or categorical biological activity of one compound based on a quantitative description of the molecular structure.<sup>18</sup> Merget and colleagues successfully employed RF to generate ligand-based prediction models for over 280 kinases, and in their work, RF generally outperforms alternative machine learning models.<sup>19</sup> Deep learning-based methods have rapidly emerged to provide state-of-the-art performance in fields such as computer vision and natural language processing.<sup>20,21</sup> Neural networks have also been successfully applied in the cheminformatic domain through creative manipulation of 2D or 3D chemical structures and construction of the network architecture.<sup>22–24</sup> Among these ML algorithms, Extreme Gradient Boosting (XGBoost) appears to be a very effective and efficient machine-learning method in the realm of QSAR. It can make predictions, on the average, better than those of Random Forest and almost as good as those of deep neural nets with much less computational effort.<sup>25</sup>

In this paper, XGBoost was employed to build JAK2-centric classification and regression models. Three groups of models were developed based on three kinds of fingerprints (FPs), and each group comprised a classification model and a regression model. These models would serve as tools to screen JAK2 inhibitors from the ZINC database.<sup>26</sup> The classification models were tested on the DUD-E set containing JAK2 inhibitors and decoys and further evaluated by comparing the performance on VS test sets with docking methods. The regression models were built to predict the specific activity values of active molecules identified by classification models. In order to minimize the scope of experimental screening and reduce the false positive rate, molecules that were simultaneously identified by three groups of models would be subjected to *in vitro* bioactivity evaluation. Here, we are dedicated to quickly and efficiently obtaining JAK2 inhibitors from a large molecular database via our approach of screening based on machine learning models.

## 2. MATERIALS AND METHODS

**2.1. Data Collection.** **2.1.1. Data Set 1.** For the work described here, we mainly collected data from PubChem and the Binding database (BindingDB). To build classification models, 4607 active molecules of JAK2 with their molecular structures and  $IC_{50}$  values and 216,460 compounds, which were found inactive against JAK2 with molecular structures, were collected from PubChem. Additionally, 6149 JAK2 inhibitors with  $IC_{50}$  values were downloaded from BindingDB.

To combine the data collected from the two databases, the CID code was chosen as the unique identification and the repeated molecules with lower activity were further removed. As a result, 7234 active and 216,460 inactive molecules of JAK2 were collected.

For developing machine learning models, the data set should be split into training and test sets. Data splitting by random division has been demonstrated to offer more realistic predictions of the learning algorithms.<sup>27</sup> Thus, we randomly split the data into a training set and a test set by a 4:1 ratio using “shuffle” in Python. The resulted training set contained 178,955 molecules (5787 active and 173,168 inactive molecules), and there were 44,739 molecules in the test set (1447 active and 43,292 inactive molecules).

**2.1.2. Data Set 2.** Based on data set 1, the molecules of JAK2 were further divided followed by a threshold rule: active ( $\leq 10 \mu\text{M}$ ) and inactive ( $> 10 \mu\text{M}$ ). The reasons for choosing a threshold of  $10 \mu\text{M}$  are as follows: (1) the  $10 \mu\text{M}$  is a cutoff for starting follow-up activities after high-throughput screening (HTS).<sup>28</sup> (2) Another is to see if the threshold rule could exert good effects on the models. Finally, there were 6734 active molecules (5387 molecules in the training set and 1347 molecules in the test set) and 216,960 inactive molecules (173,568 molecules in the training set and 43,392 molecules in the test set).

In data sets 1 and 2, active compounds were heavily outnumbered by inactive compounds, which was referred to as the class imbalance problem. The class imbalance problem could lead to poor performance of ML algorithms.<sup>29</sup> Ensemble learning methods of which Boosting and Bagging are the most successful approaches have been extensively used to handle class imbalance problems.<sup>30</sup> In this work, the classification models were developed based on XGBoost (one of the boosting approaches). The area under the receiver operating characteristic (ROC) curve (AUC), which was suitable to evaluate imbalanced data sets, was employed to evaluate the performance of our classification models.

**2.1.3. Data Set 3.** Active molecules and their  $IC_{50}$  values were collected from data set 1. Particularly, two molecules whose  $IC_{50}$  values were far greater than  $1000 \mu\text{M}$  were removed. Thus, the  $IC_{50}$  values of the data set ranged from  $0.004 \text{ nM}$  to  $1000 \mu\text{M}$ , which was wide enough to build a good activity prediction model. Finally, there were 7232 active molecules (5786 molecules in the training set and 1446 molecules in the test set).

**2.1.4. Data Set from DUD-E.**<sup>31</sup> In addition, 153 active molecules and 6500 decoys of JAK2 were collected from DUD-E to validate the performance of the classification models.

**2.1.5. Data Set of the VS Test.** Furthermore, in order to test the VS effectiveness of the classification models, 13 JAK2 inhibitors were collected from the literature, which were not included in the two databases mentioned above. The  $IC_{50}$  values and molecular structures are shown in Table S1. Moreover, we downloaded three sets of inactive molecules from ZINC, which contained 2000, 10,000, and 20,000 molecules.<sup>32–43</sup>

**2.2. Chemical Representations.** All the molecules were represented by three types of molecular FPs: (1) The MACCS fingerprint uses “1” or “0” to indicate the presence or absence of the substructure, and the length of which is 166 bits. (2) The extended-connectivity fingerprint (ECFP) is a vector with a fixed length (e.g., 1024 bits), which initially uses unique

identifiers to demonstrate structures around all heavy atoms of a molecule with a defined radius, and can be classified as ECFP\_2, ECFP\_4, ECFP\_6, etc. The appended number is the effective diameter of the largest feature and is equal to twice the number of iterations performed. For example, if three iterations are performed, the largest possible fragment will have a width of six bonds.<sup>44</sup> (3) Mol2vec is an unsupervised machine learning approach inspired by natural language processing techniques. Mol2vec learns vector representations of molecular substructures that are pointing to similar directions for chemically related substructures. Compounds can be encoded as vectors via summing up the vectors of the individual substructures and then fed into modeling approaches for prediction of compound properties. In this study, the Mol2vec model was pre-trained based on a corpus containing 19.9 million compounds and then utilized to feature new samples.<sup>45</sup> Finally, 300-dimensional embeddings were generated for all compounds. MACCS and ECFP were calculated using the RDKit<sup>46</sup> (version 2019.03.1). Here, three kinds of chemical representations were employed for the generation of classification models and regression models, herein after termed MACCS+ECFP2 (the two FPs were concatenated for each compound), Mol2vec, and ECFP\_4.

**2.3. Methods for Model Building.** In this study, we used a scalable end-to-end tree boosting system called XGBoost (<https://github.com/dmlc/xgboost>), which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges.<sup>47</sup> XGBoost builds on previous ideas in gradient boosting, which builds a sequential series of smaller trees where each tree corrects for the residuals in the predictions made by all the previous trees. XGBoost has many adjustable parameters compared with RF, which has a handful of adjustable parameters (e.g., number of trees, fraction of descriptors used at each branching, node size, etc.).<sup>25</sup> Here, we introduce some parameters we used to build classification and regression models: eta (step size shrinkage) was set to 0.1; the max depth (maximum depth of a tree) was set to 10; the colsample\_bytree (what fraction of descriptors would be examined for each tree) was set to 0.7; and the colsample\_bylevel was set to 0.7. Gamma was set to 0.1; the objective was set to “binary: logistic” and “reg: linear”.

**2.4. Cross-Validation and Model Evaluation.** The training data set were split into five equal parts for a fivefold cross-validation (CV). The performances of classification models were evaluated by the following metrics: AUC, accuracy (Q), sensitivity (SE), specificity (SP), and precision (PR). The metric used to evaluate the performance of regression models was  $R^2$ , which was the coefficient of determination between predicted and observed activities in the test set.  $R^2$  measures the degree of concordance between the predictions and corresponding observations. Here,  $R^2$  was calculated by `r2_score` available in `scikit-learn`.<sup>48</sup> Some of the metrics were calculated using the equation in Table 1.

**2.5. Cell-Free Kinase Activity Assays.** Homogeneous time-resolved fluorescence (HTRF) assays were conducted to evaluate the inhibition of JAKs by different compounds.<sup>49</sup> The assays were performed with the HTRF KinEASE kit (Cisbio Bioassays, Codolet, France) according to the manufacturer's instructions. Briefly, test compounds were diluted in DMSO with a 10-fold gradient series to generate a six-point curve with an initial concentration of 10  $\mu$ M. The enzymes were mixed with the test compounds and the peptide substrates in the kinase reaction buffer. Following the addition of related

**Table 1. Description of the Evaluation Metrics to Assess the Classification Model**

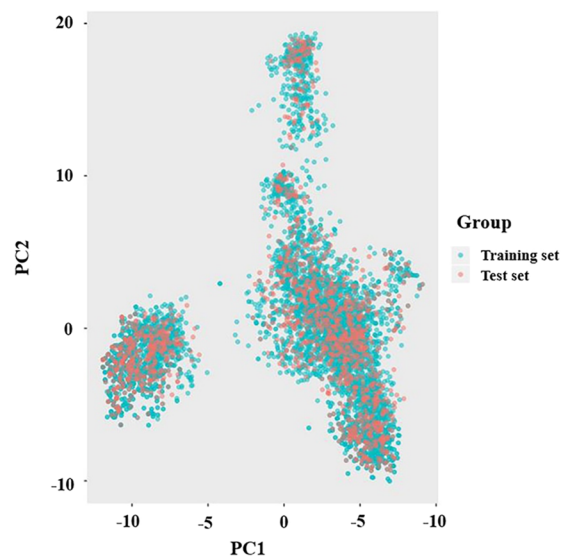
metrics	equation <sup>a</sup>
accuracy (Q)	$(TP + TN)/N$
sensitivity (SE)	$TP/(TP + FN)$
specificity (SP)	$TN/(TN + FP)$
precision (PR)	$TP/(TP + FP)$
MCC	$S = (TP + FN)/N$ $P = (TP + FP)/N$ $MCC = \frac{TP - N \cdot SP}{\sqrt{PS(1-S)(1-P)}}$

<sup>a</sup>TP is the number of correctly predicted actives (true positives), TN is the number of correctly predicted inactives (true negatives), N is the total number of molecules in the database, FN is the number of mispredicted inactives (false negatives), FP is the number of mispredicted positives (false positives). The MCC (Matthews correlation coefficient) is applied to evaluate the performance of classification models. The perfect model produces an MCC value of 1.

reagents, the signal of time-resolved fluorescence energy transfer (TR-FRET) was detected using a Synergy H1 microplate reader (BioTek Instruments, Winooski, Vermont, U.S.A.). The half maximal inhibitory concentration ( $IC_{50}$ ) was calculated by nonlinear regression.

### 3. RESULTS AND DISCUSSION

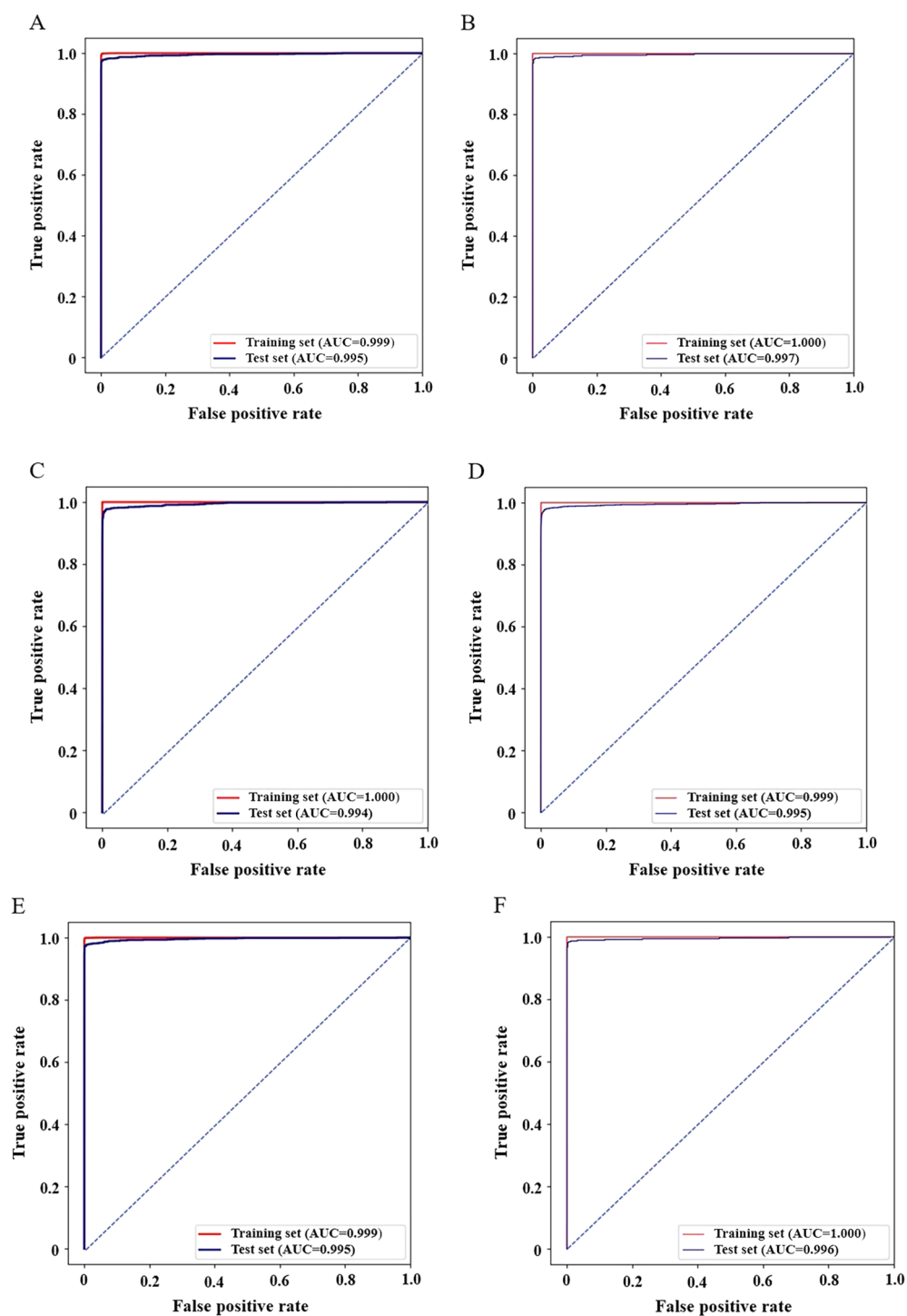
**3.1. Chemical Diversity Analysis.** To verify the diversity of chemical space of the JAK2 inhibitors that we collected to



**Figure 2.** First two principal components of PCA of JAK2 inhibitors. Besides the majority cluster, the data occupies several further distinct clusters in the chemical space.

develop the classification models and regression models, a principal component analysis (PCA)<sup>50</sup> was performed on the 7234 active molecules of JAK2 with ECFP\_4 as input. As demonstrated by the chemical space defined by the first two principal components in Figure 2, distinct clusters indicating high diversity were observed and the chemical space of the training set overlapped that of the test set.

**3.2. Classification Models.** Based on three types of fingerprints and two kinds of data sets (data set 1 and data set 2), six classification models were built (models 1A, 1B, 2A, 2B,



**Figure 3.** Receiver operating characteristic (ROC) curves of models 1A, 1B, 2A, 2B, 3A, and 3B. (A) ROC curve of model 1A. (B) ROC curve of model 1B (C) ROC curve of model 2A. (D) ROC curve of model 2B. (E) ROC curve of model 3A. (F) ROC curve of model 3B.

**Table 2.** Performance of the Six Classification Models on the Test Set

model	data set	fingerprints		training set		test set			
		type	length	5-CV(AUC)	SE	SP	PR	Q	MCC
model 1A	1	ECFP_2 + MACCS	1024 + 166	0.994	0.9481	0.9998	0.9928	0.9981	0.9692
model 1B	2	ECFP_2 + MACCS	1024 + 166	0.996	0.9636	0.9996	0.9871	0.9985	0.9745
model 2A	1	Mol2vec	300	0.994	0.9109	0.9995	0.9836	0.9966	0.9448
model 2B	2	Mol2vec	300	0.994	0.9065	0.9995	0.9823	0.9967	0.9420
model 3A	1	ECFP_4	1024	0.995	0.9488	0.9998	0.9935	0.9981	0.9696
model 3B	2	ECFP_4	1024	0.996	0.9584	0.9997	0.9885	0.9984	0.9725



**Table 3. Performance on the DUD-E Set of Six Models**

models	SE	SP	PR	Q	MCC
model 1A	0.8824	0.9995	0.9782	0.9969	0.9275
model 1B	0.8954	0.9998	0.9928	0.9975	0.9416
model 2A	0.8824	0.9982	0.9184	0.9956	0.8979
model 2B	0.8758	0.9983	0.9241	0.9955	0.8974
model 3A	0.7582	0.9998	0.9915	0.9944	0.8645
model 3B	0.9216	0.9998	0.9930	0.9981	0.9556

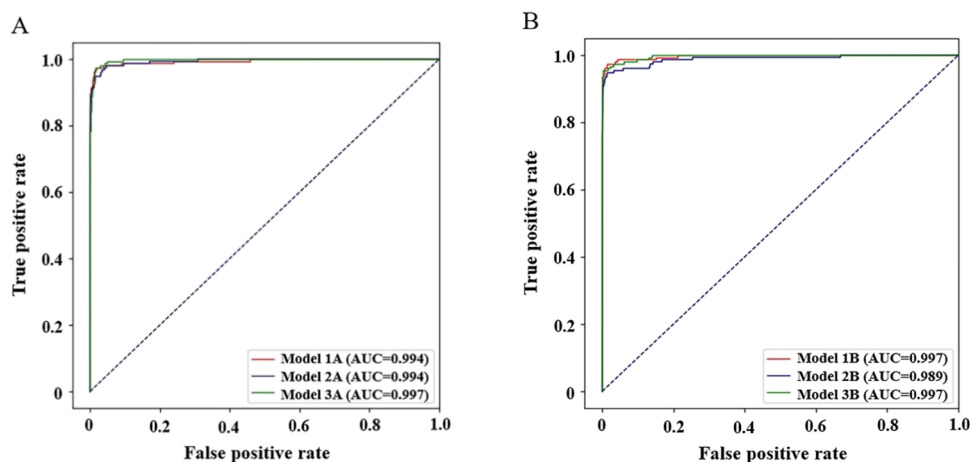
3A, and 3B). The models were tested on the internal test sets and further evaluated on the DUD-E data set and VS test set to evaluate generalization ability and VS effectiveness of our models.

**3.2.1. Performance on the Internal Test Sets.** The AUC values of our models are shown in Figure 3, which were all close to 1, showing great predictive power of our classification models. The detailed prediction performance of the six models is shown in Table 2. In general, our six models have a high capacity to separate actives from inactives in terms of SE, SP, PR, Q, and MCC values, which all exceeded 0.9. The performances of models 1B and 3B that were built based on data set 2 were better than those of models 1A and 3A that were based on data set 1 (without a threshold rule) in terms of MCC values. The performances of model 2B were nearly the same as those of model 2A, and performances of models that were built on Mol2vec were not as good as those on the other two types of FPs, suggesting that Mol2vec was different with the other two and was not sensitive to the 10  $\mu$ M threshold applied to data set 2. The SE values of six models were approximately 0.9 and were not good enough compared to the SP values, which were close to 1. Sensitivity (also known as the recall), which measures the proportion of positives that are correctly identified, may be related to the number of active molecules fed into the model. Similarly, specificity (also called as the true negative rate) measures the proportion of negatives that are correctly identified and may have a close relationship with the number of inactive molecules. Since the number of active molecules on the training set was far less than the inactive molecules, the SE values of our models were not high enough compared to the SP values.

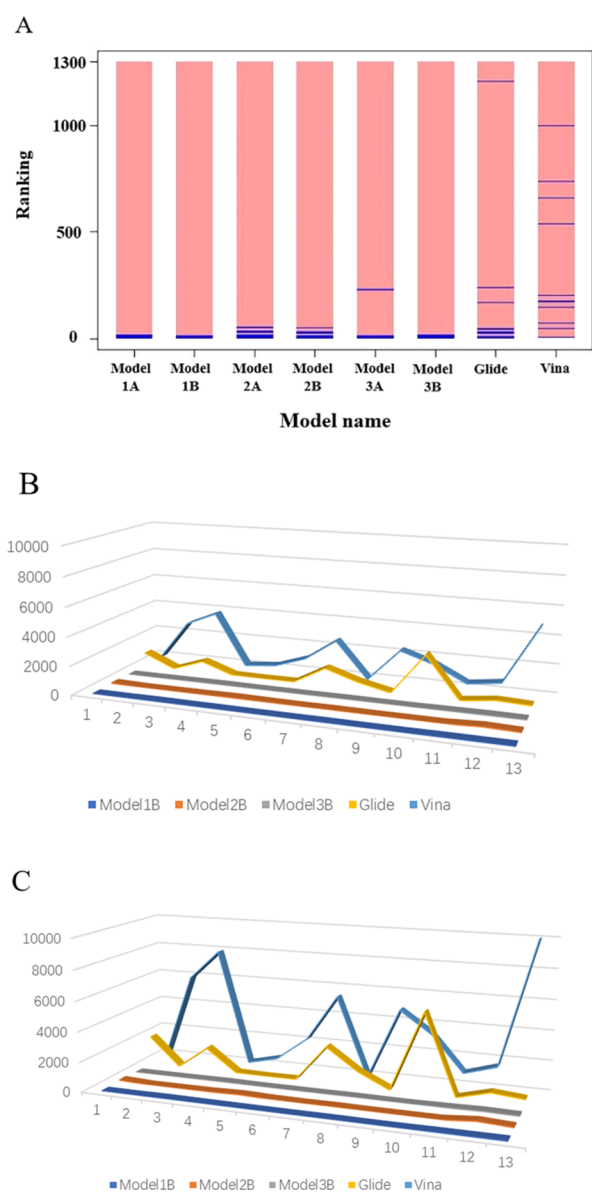
**3.2.2. Prediction on the DUD-E Data Set.** In order to assess the generalization ability of the classification models and

further compare the performance of models of A with the models of B, we tested our models on external JAK2 molecular data collected from the DUD-E data set. Being consistent with the results of internal evaluation, the performance of models 1B and 3B was greater than that of models 1A and 3A, and there were no distinct differences between model 2A and model 2B in terms of the metric values (Table 3). Although the SE and MCC values exhibited a slight drop compared to the results of internal test sets, they were high enough with an average value close to 0.9, and the values of other metrics were all equally great with those of the internal test. We could conclude that our models have great generalization ability. The AUC values were also analyzed (Figure 4).

**3.2.3. Prediction on the VS Test Set.** Glide and AutoDock Vina were employed to evaluate the performance of our classification models.<sup>51</sup> Details of the two docking methods are shown in Table S2. The performance on the test set of 2013 molecules of the six models were significantly better than that of the two docking methods (Figure 5A). Models 1A, 1B, and 3B exhibited superior enrichment power that ranked 13 active molecules in the top 15 (0.1% false positive rate), and model 1B was the best model that ranked 13 active molecules in the top 13. Models 2A and 2B ranked 12 actives in the top 33 (1% false positive rate); the last one was ranked at 50. Obviously, model 3B had a greater performance than that of model 3A for recognizing an active compound that was ranked at 229 by the latter. Instead, Glide could rank five actives in the top 15 and six in the top 33. AutoDock Vina ranked one active molecule in the top 15 and top 33. The two structure-based docking methods ranked molecules by scores calculated by their inner methods. However, there was no standard of scores to discriminate actives and inactives, which makes it possible that the top 1 was the inactive compound when the data set comprised inactives. As for classification models, compounds were ranked according to the predicted class-membership probability values, which generally had a threshold of 0.5. Compounds with probability values greater than 0.5 were denominated as actives and those otherwise as inactives. In general, the performances of models of B were better than those of models of A on the VS test, and the former were further compared with the two docking methods on two other test sets containing 10,013 and 20,013 compounds. The increase in the number of molecules was inevitably



**Figure 4.** ROC curves of models 1A, 2A, 3A, 1B, 2B, and 3B. (A) ROC curves of models 1A, 2A, and 3A. (B) ROC curves of models 1B, 2B, and 3B.



**Figure 5.** (A) Comparison of the ability to distinguish actives and inactives of models 1A, 1B, 2A, 2B, 3A, and 3B, Glide, and AutoDock Vina on the VS test set (2013 molecules). Active molecules (blue) are expected to gather at the bottom of the histogram. (B) Comparison of the performance of models 1B, 2B, 3B, Glide, and AutoDock Vina on the VS test (10,013 molecules). (C) Comparison of the performance of models 1B, 2B, and 3B, Glide, and AutoDock Vina on the VS test (20,013 molecules).

accompanied by an increase in calculation time. However, it took about 2 min to calculate 10,000 molecules by our classification models. In contrast, AutoDock Vina took approximately 70 h to calculate 10,000 molecules, and the speed of Glide (standard precision) was nearly the same. With the increase of interference molecules, the sharp drop of the enrichment power of the two docking methods could be clearly observed (Figure 5B,C and Table S3).

Combining the VS test results with the results of the internal test and DUD-E test, we could conclude that our six classification models were all excellent in the identification of JAK2 inhibitors and the threshold rule improved the performance of models 1A and 3A, whereas there was no

distinct difference between models 2B and 2A. Thus, models 1B, 2B, and 3B were chosen as the final classification models in the three groups.

**3.3. Regression Models.** Based on three different types of chemical representations and data set 3, we built three regression models (models 1C, 2C, and 3C) to predict  $IC_{50}$  values of the active molecules.  $R^2$  of three models calculated on the training sets were 0.97, 0.97, and 0.95 and on the test sets were 0.80, 0.78, and 0.80 (Figure 6). The performances of our regression models were in line with those of SVR models in the current study.<sup>52</sup>

To further evaluate the generalization ability of our regression models, JAK2 inhibitors that had a smile format of molecules and exact  $IC_{50}$  values were collected from the ChEMBL database. Since the 10  $\mu$ M threshold rule was applied in the classification model and our regression model was developed to predict the activity of active molecules, we further removed molecules whose  $IC_{50}$  values were more than 10  $\mu$ M and got 4116 molecules from ChEMBL. The  $R^2$  of three models on the external test set were 0.80, 0.78, and 0.78 (Figure 7).

**3.4. Virtual Screening and Biological Evaluation.** To accurately and efficiently acquire JAK2 inhibitors from the ZINC database, we developed a hierarchical strategy to integrate three classification models and corresponding three activity prediction models (Figure 8). Generally, our hierarchical procedure consists of three steps: (1) filtering the raw data set using Lipinski's rule of five, (2) classifying the molecules that passed the structure filter into actives and inactives by classification models and predicting the activity values of the actives using regression models, and (3) picking molecules according to their predicted  $IC_{50}$  values and chemical structure features.

In detail, RDKit was used to calculate properties of the 7234 JAK2 inhibitors, such as the molecular weight,  $A \log P$ , and the number of hydrogen-bond donors and acceptors. Thus, there was a range for their properties that could be used to roughly filter the ZINC database containing 42,271,452 molecules (downloaded from <http://zinc12.docking.org/subsets/everything>). A total of 20,761,052 compounds passed the filter and were subjected to models 1B, 2B, and 3B. These three classification models predicted the relationship possibilities that ranged from 0 to 1 of the input molecules, and the value of 1 meant that this model considered the molecule to have a 100% probability of being active. Here, 18,100, 56,144, and 18,082 molecules whose values of probability were greater than 0.5 were retrieved from three classification models and were put into three corresponding regression models. To acquire highly potent JAK2 inhibitors, we retained 5537, 9758, and 5809 molecules whose predicted activity values were greater than 100 nM from the three regression models. To make the final molecules simultaneously satisfy the three groups of models according to their CID number, 1702 molecules that repeatedly existed in the results of the three activity prediction models were selected. Considering that the pyrrolopyrimidine scaffold was used by three FDA-approved drugs, we decided to pick molecules with the same or similar scaffold (e.g., pyrrolopyridine) to conduct the experimental evaluation. Thus, 30 molecules with a pyrrolopyrimidine-like scaffold were picked, and 13 molecules, which were commercially available, were subjected to in vitro biological evaluation.

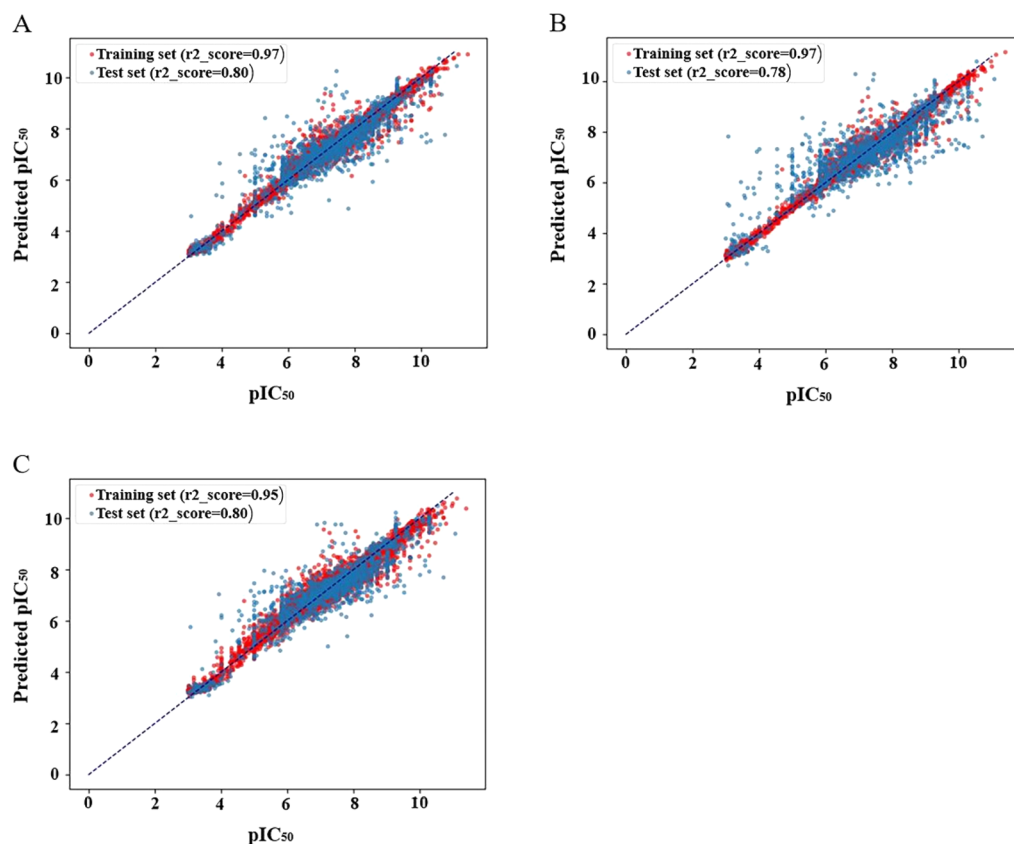


Figure 6. Scatter plot of models (A) 1C, (B) 2C, and (C) 3C on the training and test sets.

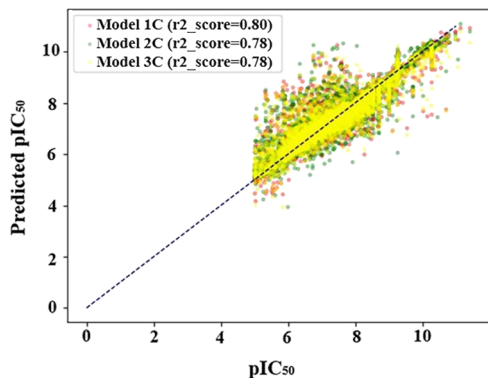


Figure 7. Scatter plot of models 1C, 2C, and 3C on the external test set.

The results of biological evaluation are shown in Table 4. (The details for structure confirmation of 13 purchased compounds are shown in Figure S3.) Ruxolitinib and tofacitinib were employed as positive drugs for JAK2 and JAK3 inhibitors, respectively. Ruxolitinib showed high JAK2 potency ( $JAK2 IC_{50} < 1 \text{ nM}$ ) and tofacitinib showed high JAK3 potency ( $JAK3 IC_{50} < 1 \text{ nM}$ ). Among 13 biological tested compounds, eight compounds showed potency against JAK2, and the  $IC_{50}$  values of six compounds were identified to be less than 100 nM. Compound 9 exhibited high JAK2 potency ( $JAK2 IC_{50} < 1 \text{ nM}$ ) and high selectivity versus JAK3 ( $IC_{50} = 694 \text{ nM}$ ). The predicted and actual biological values of the 13 compounds were compared, and some showed high accuracy (Figure 9). The inhibitory profiles of compounds that showed potency against JAK2 or JAK3 clearly showed a dose-

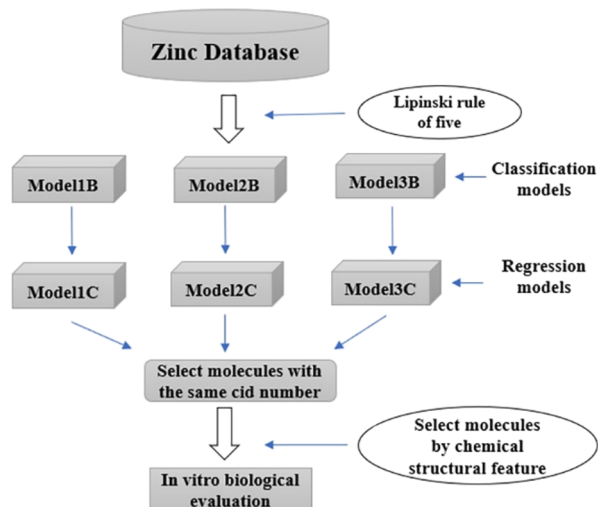
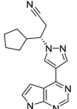
















Figure 8. Flowchart of our ligand-based hierarchical screening strategy.

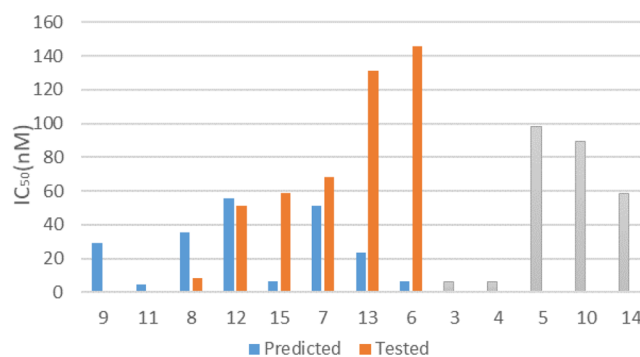
dependent pattern (Figure S1). PAINS screening was applied to these compounds, and all of them passed the test<sup>53</sup> (Figure S2). Though all 13 molecules used pyrrolopyrimidine-like scaffolds, five compounds did not show potency against JAK2, and the  $IC_{50}$  values of eight compounds spanned at least 4 orders of magnitude, indicating that, except for the scaffold, the structures of other components of the compounds were important for potency and selectivity. Compounds 9 and 11 were docked into JAK2 (PDB ID: 2XA4) to see the binding mode of molecules that exhibited high potency against JAK2 (Figure 10). These two compounds interact with a backbone

**Table 4. Active Compounds Identified by Models and the Results of Their in Vitro Biological Tests<sup>a</sup>**

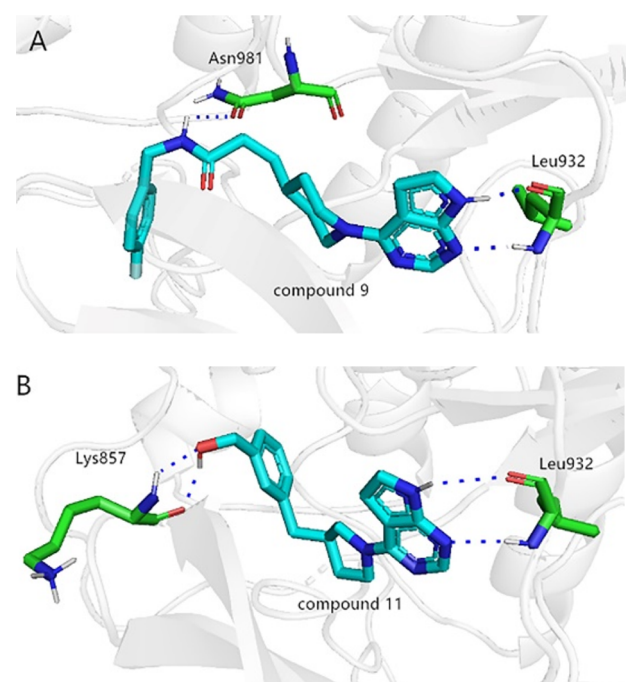
number	Structure	IC <sub>50</sub> ( $\mu$ M)		Predicted IC <sub>50</sub> ( $\mu$ M)
		JAK2	JAK3	
1		<0.001	- <sup>a</sup>	0.0018
2		-	<0.001	0.0049
3		-	-	0.0059
4		-	-	0.0061
5		-	-	0.0981
6		1.455	-	0.0062
7		0.0682	-	0.0509
8		0.0085	0.3779	0.0356
9		<0.001	0.6940	0.0289
10		-	-	0.0894
11		<0.001	-	0.0044
12		0.0511	>1	0.0553
13		0.1315	0.0349	0.0232
14		-	-	0.0585
15		0.0589	-	0.0065

<sup>a</sup>No significant inhibitory effects were observed.

of hinge residues (Leu932) by a pyrrolopyrimidine scaffold, and other parts of the two molecules interact with different residues of the protein. These 13 molecules would provide us with diverse chemical structures connected to the same



**Figure 9.** Column chart of predicted and tested IC<sub>50</sub> values. Compounds that were identified as inactive are colored gray.



**Figure 10.** (A) Binding mode of compound 9 (cyan stick) in the context of JAK2 (white cartoon). The hydrogen bonds between compound 9 and residues Leu932 and Asn981 are illustrated as blue lines. (B) Binding mode of compound 11 (cyan stick) in the context of JAK2 (white cartoon). The hydrogen bonds between compound 11 and residues Leu932 and Lys857 are illustrated as blue lines.

scaffold for designing and optimizing molecules based on the pyrrolopyrimidine scaffold to improve their potency on JAK2 and selectivity versus other JAKs.

#### 4. CONCLUSIONS

In the present work, the development and application of classification models and regression models based on XGBoost methods were reported. We developed six classification models based on three different types of FPs and a threshold of splitting in the data set and evaluated the six models by comparing their performances on internal test sets, the DUD-E set and VS test set. The results showed that employing 10  $\mu$ M as the threshold of JAK2 inhibitors and applying this rule to the data set enhanced the quality of the models. The best three models produced MCC values of 0.94, 0.97, and 0.94. We also built three regression models based on the three FPs, and their  $R^2$  values calculated for test sets were 0.80, 0.78, and 0.80. We



could see that models 1B and 3B performed equally on all kinds of tests, and the  $R^2$  values of regression models 1C and 3C were almost equal, which might be because they all used ECFP fingerprints to represent their molecular structures. The quality of models based on Mol2vec was slightly inferior to that based on ECFP, and the distinct differences were reflected on the number of molecules that passed the classification and regression models. From the screening results, it could be seen that the intersection of the three groups could at least reduce the number of molecules by two-thirds. Finally, we selected 13 commercially available compounds with pyrrolopyrimidine-like scaffolds to conduct the experimental evaluation, and eight of them showed activity against JAK2. The  $IC_{50}$  values of six compounds were identified to be less than 100 nM. Compounds **9** and **11** exhibited high JAK2 potency and high selectivity versus JAK3. These 13 molecules would provide ideas for our subsequent pyrrolopyrimidine-based molecular optimization. We expect that our strategy may be generally applicable in ligand-based campaigns and our current work may serve as a starting point to develop novel JAK2 inhibitors with high potency and selectivity.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.9b00798>.

Detailed information of 13 JAK2 inhibitor compounds with experimental bioactivities, values, and structure; detailed information of the two docking methods; and detailed information of ranking of 13 JAK2 inhibitors in the three VS test sets (Tables S1–S3) (PDFs)

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [wangxiaojian@imm.ac.cn](mailto:wangxiaojian@imm.ac.cn).

### ORCID

Xiaojian Wang: 0000-0002-1856-8820

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was financially supported by the CAMS Collaborative Innovation Project (nos. 2017-I2M-3-011 and 2019-I2M-1-005) and 2018 Youth Talent Award Program for the Basic Scientific Research of the Medical College (no. 2018RC350009). We thank Dr. Yadong Chen for the support of the docking studies.

## ■ REFERENCES

- (1) Leonard, W. J.; O'Shea, J. J. Jaks and Stats: Biological Implications. *Annu. Rev. Immunol.* **1998**, *16*, 293–322.
- (2) Wilks, A. F. The Jak Kinases: Not Just Another Kinase Drug Discovery Target. *Semin. Cell Dev. Biol.* **2008**, *19*, 319–328.
- (3) Quintas-Cardama, A.; Vaddi, K.; Liu, P.; Manshour, T.; Li, J.; Scherle, P. A.; Caulder, E.; Wen, X.; Li, Y.; Waeltz, P.; Rupa, M.; Burn, T.; Lo, Y.; Kelley, J.; Covington, M.; Shepard, S.; Rodgers, J. D.; Haley, P.; Kantarjian, H.; Fridman, J. S.; Verstovsek, S. Preclinical Characterization of the Selective Jak1/2 Inhibitor Incb018424: Therapeutic Implications for the Treatment of Myeloproliferative Neoplasms. *Blood* **2010**, *115*, 3109–3117.
- (4) Mesa, R. A.; Yasothan, U.; Kirkpatrick, P. Ruxolitinib. *Nat. Rev. Drug Discov.* **2012**, *11*, 103–104.

- (5) Flanagan, M. E.; Blumenkopf, T. A.; Brissette, W. H.; Brown, M. F.; Casavant, J. M.; Shang-Poa, C.; Doty, J. L.; Elliott, E. A.; Fisher, M. B.; Hines, M.; Kent, C.; Kudlacz, E. M.; Lillie, B. M.; Magnuson, K. S.; McCurdy, S. P.; Munchhof, M. J.; Perry, B. D.; Sawyer, P. S.; Strelevitz, T. J.; Subramanyam, C.; Sun, J.; Whipple, D. A.; Changelian, P. S. Discovery of Cp-690,550: A Potent and Selective Janus Kinase (Jak) Inhibitor for the Treatment of Autoimmune Diseases and Organ Transplant Rejection. *J. Med. Chem.* **2010**, *53*, 8468–8484.

- (6) Fridman, J. S.; Scherle, P. A.; Collins, R.; Burn, T. C.; Li, Y.; Li, J.; Covington, M. B.; Thomas, B.; Collier, P.; Favata, M. F.; Wen, X.; Shi, J.; McGee, R.; Haley, P. J.; Shepard, S.; Rodgers, J. D.; Yeleswaram, S.; Hollis, G.; Newton, R. C.; Metcalf, B.; Friedman, S. M.; Vaddi, K. Selective Inhibition of Jak1 and Jak2 Is Efficacious in Rodent Models of Arthritis: Preclinical Characterization of Incb028050. *J. Immunol.* **2010**, *184*, 5298–5307.

- (7) Baxter, E. J.; Scott, L. M.; Campbell, P. J.; East, C.; Fourouclas, N.; Swanton, S.; Vassiliou, G. S.; Bench, A. J.; Boyd, E. M.; Curtin, N.; Scott, M. A.; Erber, W. N.; Green, A. R. Acquired Mutation of the Tyrosine Kinase Jak2 in Human Myeloproliferative Disorders. *Lancet.* **2005**, *365*, 1054–1061.

- (8) Levine, R. L.; Wadleigh, M.; Cools, J.; Ebert, B. L.; Wernig, G.; Huntly, B. J.; Boggon, T. J.; Wlodarska, I.; Clark, J. J.; Moore, S.; Adelsperger, J.; Koo, S.; Lee, J. C.; Gabriel, S.; Mercher, T.; D'Andrea, A.; Frohling, S.; Dohner, K.; Marynen, P.; Vandenberghe, P.; Mesa, R. A.; Tefferi, A.; Griffin, J. D.; Eck, M. J.; Sellers, W. R.; Meyerson, M.; Golub, T. R.; Lee, S. J.; Gilliland, D. G. Activating Mutation in the Tyrosine Kinase Jak2 in Polycythemia Vera, Essential Thrombocythemia, and Myeloid Metaplasia with Myelofibrosis. *Cancer Cell* **2005**, *7*, 387–397.

- (9) James, C.; Ugo, V.; Couédic, J.-P. L.; Staerk, J.; Delhommeau, F.; Lacout, C.; Garçon, L.; Raslova, H.; Berger, R.; Bennaceur-Griscelli, A.; Villeval, J. L.; Constantinescu, S. N.; Casadevall, N.; Vainchenker, W. A Unique Clonal Jak2 Mutation Leading to Constitutive Signalling Causes Polycythaemia Vera. *Nature* **2005**, *434*, 1144–1148.

- (10) Kralovics, R.; Passamonti, F.; Buser, A. S.; Teo, S. S.; Tiedt, R.; Passweg, J. R.; Tichelli, A.; Cazzola, M.; Skoda, R. C. A Gain-of-Function Mutation of Jak2 in Myeloproliferative Disorders. *N. Engl. J. Med.* **2005**, *352*, 1779–1790.

- (11) Menet, C. J.; Rompaey, L. V.; Geney, R. Advances in the Discovery of Selective Jak Inhibitors. *Progr. Med. Chem.* **2013**, *52*, 153–223.

- (12) Jasuja, H.; Chadha, N.; Kaur, M.; Silakari, O. Dual Inhibitors of Janus Kinase 2 and 3 (Jak2/3): Designing by Pharmacophore- and Docking-Based Virtual Screening Approach. *Mol. Diversity* **2014**, *18*, 253–267.

- (13) Singh, K. D.; Karthikeyan, M.; Kirubakaran, P.; Nagamani, S. Pharmacophore Filtering and 3d-Qsar in the Discovery of New Jak2 Inhibitors. *J. Mol. Graphics Modell.* **2011**, *30*, 186–197.

- (14) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

- (15) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

- (16) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.

- (17) Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W. Svm Model for Virtual Screening of Lck Inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 877–885.

- (18) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

- (19) Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J. Med. Chem.* **2017**, *60*, 474–485.

- (20) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. Acm.* **2017**, *60*, 84–90.

- (21) Goldberg, Y. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420.

- (22) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G.  $K_{DEEP}$ : Protein-Ligand Absolute Binding Affinity Prediction Via 3d-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296.
- (23) Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J. Chem. Inf. Model.* **2019**, *59*, 1073–1084.
- (24) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.
- (25) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360.
- (26) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. Zinc: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
- (27) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of Qsar Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570–2578.
- (28) Briem, H.; Gunther, J. Classifying “Kinase Inhibitor-Likeness” by Using Machine-Learning Methods. *ChemBioChem* **2005**, *6*, 558–566.
- (29) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. Qsar Modeling of Imbalanced High-Throughput Screening Data in Pubchem. *J. Chem. Inf. Model.* **2014**, *54*, 705–712.
- (30) Guo, X.; Yin, Y.; Dong, C.; Yang, G.; Zhou, G., On the Class Imbalance Problem. In *Fourth International Conference on Natural Computation*, 2008, 2008; pp 192–201.
- (31) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (Dud-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
- (32) Abdel-Magid, A. F. Janus-Associated Kinase 1 (Jak1) Inhibitors as Potential Treatment for Immune Disorders. *ACS Med. Chem. Lett.* **2017**, *8*, 598–600.
- (33) Casimiro-Garcia, A.; Trujillo, J. I.; Vajdos, F.; Juba, B.; Banker, M. E.; Aulabaugh, A.; Balbo, P.; Bauman, J.; Chrencik, J.; Coe, J. W.; Czerwinski, R.; Dowty, M.; Knafels, J. D.; Kwon, S.; Leung, L.; Liang, S.; Robinson, R. P.; Telliez, J. B.; Unwalla, R.; Yang, X.; Thorarensen, A. Identification of Cyanamide-Based Janus Kinase 3 (Jak3) Covalent Inhibitors. *J. Med. Chem.* **2018**, *61*, 10665–10699.
- (34) Elsayed, M. S. A.; Nielsen, J. J.; Park, S.; Park, J.; Liu, Q.; Kim, C. H.; Pommier, Y.; Agama, K.; Low, P. S.; Cushman, M. Application of Sequential Palladium Catalysis for the Discovery of Janus Kinase Inhibitors in the Benzo[*C*]Pyrrolo[2,3-*H*][1,6]Naphthyridin-5-One (Bpn) Series. *J. Med. Chem.* **2018**, *61*, 10440–10462.
- (35) Grimster, N. P.; Anderson, E.; Alimzhanov, M.; Bebernitz, G.; Bell, K.; Chuaqui, C.; Deegan, T.; Ferguson, A. D.; Gero, T.; Harsch, A.; Huszar, D.; Kawatkar, A.; Kettle, J. G.; Lyne, P.; Read, J. A.; Rivard Costa, C.; Ruston, L.; Schroeder, P.; Shi, J.; Su, Q.; Throner, S.; Toader, D.; Vasbinder, M.; Woessner, R.; Wang, H.; Wu, A.; Ye, M.; Zheng, W.; Zinda, M. Discovery and Optimization of a Novel Series of Highly Selective Jak1 Kinase Inhibitors. *J. Med. Chem.* **2018**, *61*, 5235–5244.
- (36) Huang, Y.; Dong, G.; Li, H.; Liu, N.; Zhang, W.; Sheng, C. Discovery of Janus Kinase 2 (Jak2) and Histone Deacetylase (Hdac) Dual Inhibitors as a Novel Strategy for the Combinational Treatment of Leukemia and Invasive Fungal Infections. *J. Med. Chem.* **2018**, *61*, 6056–6074.
- (37) Jones, P.; Storer, R. I.; Sabnis, Y. A.; Wakenhut, F. M.; Whitlock, G. A.; England, K. S.; Mukaiyama, T.; Dehnhardt, C. M.; Coe, J. W.; Kortum, S. W.; Chrencik, J. E.; Brown, D. G.; Jones, R. M.; Murphy, J. R.; Yeoh, T.; Morgan, P.; Kilty, I. Design and Synthesis of a Pan-Janus Kinase Inhibitor Clinical Candidate (Pf-06263276) Suitable for Inhaled and Topical Delivery for the Treatment of Inflammatory Diseases of the Lungs and Skin. *J. Med. Chem.* **2017**, *60*, 767–786.
- (38) Kulagowski, J. J.; Blair, W.; Bull, R. J.; Chang, C.; Deshmukh, G.; Dyke, H. J.; Eigenbrot, C.; Ghilardi, N.; Gibbons, P.; Harrison, T. K.; Hewitt, P. R.; Liimatta, M.; Hurley, C. A.; Johnson, A.; Johnson, T.; Kenny, J. R.; Bir Kohli, P.; Maxey, R. J.; Mendonca, R.; Mortara, K.; Murray, J.; Narukulla, R.; Shia, S.; Steffek, M.; Ubhayakar, S.; Ultsch, M.; van Abbema, A.; Ward, S. I.; Waszkowycz, B.; Zak, M. Identification of Imidazo-Pyrrolopyridines as Novel and Potent Jak1 Inhibitors. *J. Med. Chem.* **2012**, *55*, 5901–5921.
- (39) Liang, X.; Zang, J.; Zhu, M.; Gao, Q.; Wang, B.; Xu, W.; Zhang, Y. Design, Synthesis, and Antitumor Evaluation of 4-Amino-(1*h*)-Pyrazole Derivatives as Jaks Inhibitors. *ACS Med. Chem. Lett.* **2016**, *7*, 950–955.
- (40) Ritzen, A.; Sorensen, M. D.; Dack, K. N.; Greve, D. R.; Jerre, A.; Carnerup, M. A.; Rytved, K. A.; Bagger-Bahnsen, J. Fragment-Based Discovery of 6-Arylindazole Jak Inhibitors. *ACS Med. Chem. Lett.* **2016**, *7*, 641–646.
- (41) Siu, T.; Brubaker, J.; Fuller, P.; Torres, L.; Zeng, H.; Close, J.; Mampreian, D. M.; Shi, F.; Liu, D.; Fradera, X.; Johnson, K.; Bays, N.; Kadic, E.; He, F.; Goldenblatt, P.; Shaffer, L.; Patel, S. B.; Lesburg, C. A.; Alpert, C.; Dorosh, L.; Deshmukh, S. V.; Yu, H.; Klappenbach, J.; Elwood, F.; Dinsmore, C. J.; Fernandez, R.; Moy, L.; Young, J. R. The Discovery of 3-((4-Chloro-3-Methoxyphenyl)Amino)-1-((3*r*,4*s*)-4-Cyanotetrahydro-2*h*-Pyran-3-yl)-1*H*-Pyrazole-4-Carboxamide, a Highly Ligand Efficient and Efficacious Janus Kinase 1 Selective Inhibitor with Favorable Pharmacokinetic Properties. *J. Med. Chem.* **2017**, *60*, 9676–9690.
- (42) Vazquez, M. L.; Kaila, N.; Strohbach, J. W.; Trzupek, J. D.; Brown, M. F.; Flanagan, M. E.; Mitton-Fry, M. J.; Johnson, T. A.; TenBrink, R. E.; Arnold, E. P.; Basak, A.; Heasley, S. E.; Kwon, S.; Langille, J.; Parikh, M. D.; Griffin, S. H.; Casavant, J. M.; Duclos, B. A.; Fenwick, A. E.; Harris, T. M.; Han, S.; Caspers, N.; Dowty, M. E.; Yang, X.; Banker, M. E.; Hegen, M.; Symanowicz, P. T.; Li, L.; Wang, L.; Lin, T. H.; Jussif, J.; Clark, J. D.; Telliez, J. B.; Robinson, R. P.; Unwalla, R. Identification of N-[Cis-3-[Methyl(7*h*-Pyrrolo[2,3-*D*]-Pyrimidin-4-yl)Amino]Cyclobutyl]Propane-1-Sulfo Namide (Pf-04965842): A Selective Jak1 Clinical Candidate for the Treatment of Autoimmune Diseases. *J. Med. Chem.* **2018**, *61*, 1130–1152.
- (43) Zak, M.; Mendonca, R.; Balazs, M.; Barrett, K.; Bergeron, P.; Blair, W. S.; Chang, C.; Deshmukh, G.; Devoss, J.; Dragovich, P. S.; Eigenbrot, C.; Ghilardi, N.; Gibbons, P.; Grادل, S.; Hamman, C.; Hanan, E. J.; Harstad, E.; Hewitt, P. R.; Hurley, C. A.; Jin, T.; Johnson, A.; Johnson, T.; Kenny, J. R.; Koehler, M. F.; Bir Kohli, P.; Kulagowski, J. J.; Labadie, S.; Liao, J.; Liimatta, M.; Lin, Z.; Lupardus, P. J.; Maxey, R. J.; Murray, J. M.; Pulk, R.; Rodriguez, M.; Savage, S.; Shia, S.; Steffek, M.; Ubhayakar, S.; Ultsch, M.; van Abbema, A.; Ward, S. I.; Xiao, L.; Xiao, Y. Discovery and Optimization of C-2 Methyl Imidazopyrrolopyridines as Potent and Orally Bioavailable Jak1 Inhibitors with Selectivity over Jak2. *J. Med. Chem.* **2012**, *55*, 6176–6193.
- (44) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, 742–754.
- (45) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35.
- (46) Rdkit: Open-Source Chmeinformatics Software, 2016; [Http://www.rdkit.org](http://www.rdkit.org), (accessed April 15, 2019)
- (47) Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. *arXiv:1603.02754v3*. 2016.
- (48) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (49) Harbert, C.; Marshall, J.; Soh, S.; Steger, K. Development of a Htrf @ Kinase Assay for Determination of Syk Activity. *Curr. Chem. Genomics.* **2008**, *1*, 20–26.
- (50) Ma, S.; Dai, Y. Principal Component Analysis Based Methods in Bioinformatics Studies. *Brief. Bioinform.* **2011**, *12*, 714–722.
- (51) Trott, O.; Olson, A. J. Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient

Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

(52) Miyao, T.; Funatsu, K.; Bajorath, J. Exploring Alternative Strategies for the Identification of Potent Compounds Using Support Vector Machine and Regression Modeling. *J. Chem. Inf. Model.* **2019**, *59*, 983–992.

(53) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.